**How Reliably Can Examiners Make Form Quality (FQ) Judgments in the Absence of the Form Quality (FQ) Tables?**

Rorschachiana (2021), 42(1), 21–34
https://doi.org/10.1027/1192-5604/a000135



# How reliably can examiners make FQ judgments in the absence of the FQ tables?

Dr. Claudia Pignolo[1], Dr. Donald J. Viglione[2], & Dr. Luciano Giromini[1]

[1] Department of Psychology, University of Turin, Turin, Italy

[2] California School of Professional Psychology, Alliant International University, San Diego, CA, USA

**ORCID of the authors:**

Claudia Pignolo https://orcid.org/0000-0002-6977-224X

Donald J. Viglione https://orcid.org/0000-0002-6460-1986

Luciano Giromini https://orcid.org/0000-0002-9540-4803

Donald J. Viglione receives royalties on the sale of the R-PAS manual and associated products.

Correspondence concerning this article should be addressed to Claudia Pignolo, Department of Psychology, University of Turin, Via Giuseppe Verdi 10, 10124 Torino, TO, Italy. E-mail claudia.pignolo@unito.it

**Abstract**

Form Quality (FQ) scores are well-validated measures of accuracy perceptive processes, reality testing, and severity of psychological disturbance. Research studies reveal that inter-rater reliability of FQ scoring is good when visualized objects are available in FQ tables. However, many visualized objects are not found in the FQ tables so that scoring must rely on one's individual judgment. Thus, a major question remains unsolved: how reliably can examiners make FQ judgments in the absence of the FQ tables? To address this question, we used the Rorschach Performance Assessment System (R-PAS) method. We asked 21 graduate students from our research labs to rate Form Accuracy (FA) and FQ for 86 objects from a subset of four Rorschach card (I, III, VI, and VIII). The results clearly reveal that individual examiner making FA judgements without using the FQ tables are not reliable. When scoring FQ, one should carefully scrutinize the empirically supported FQ tables and base the FQ score on these rather than personal judgements.

Keywords: Form Accuracy (FA); Form Quality (FQ); R-PAS; Rorschach

# How reliably can examiners make FQ judgments in the absence of the FQ tables?

Form Quality (FQ) is an essential variable that has been recognized for its importance since the development of the Rorschach Inkblot test (Rorschach, 1921) and refers to the "goodness of fit" of objects[1] involved in a response to the area of the blot used by the examinee. In other words, whether the object or image seen by the respondent looks like the area where it is seen in the blot. Exner (1974 , 2003), while developing the Comprehensive System (CS), identified four types of FQ: (1) *Superior-overelaborated (+),* unusually well-articulated form responses; (2) *Ordinary (o),* a high frequency response in which an object fits the blot contours; (3) *Unusual (u),* an uncommon response in which the blot contours are appropriate; (4) *Minus (-),* are of two types: Responses reported usually with low frequencies that are not congruent with the contours of the blot, and those which involve creating contours that do not exist in the blot, often called "arbitrary lines." FQ was not assigned to responses without any structure. To establish the thresholds between FQo and FQu, Exner utilized the frequency distribution of 7,500 protocols (162,427 responses), so that objects that were reported in at least 2% (150 or more) of the records in whole (W) or detail (D) areas or by at least 50 subjects in unusual detail (Dd) areas were coded as FQo, and objects with lower frequencies were coded as FQu.

Subsequently, the authors of the Rorschach Performance Assessment System (R-PAS; Meyer et al., 2011), by using a specific algorithm, combined three different sources of data to determine the R-PAS FQ codes: (1) fit, which refers to the degree to which objects reported in a specific area fit to the blot contours, (2) frequency, which refers to how often objects has been spontaneously reported by examinees at that location, and (3) the FQ coding retrieved

[1] In this paper we used to word "object" to refer to images seen by respondents and the word "entry" to refer to the words listed in the table.

from the most recent CS Tables. Thus, FQ is operationally defined as the degree to which the reported objects are common and fit the blot area. Moreover, objects were classified as ordinary (FQo), unusual (FQu), and distorted (FQ-), and responses without any structure were classified as 'none' (FQn). Overall, the R-PAS FQ tables have about 34.3% of minus (FQ-), 45.2% of unusual (FQu), and 20.5% of ordinary (FQo) objects.

FQ scores are a well-validated measure of perception accuracy, reality testing, and severity of psychological disturbance (e.g., Meyer et al. 2011; Mihura et al., 2013; Su et al., 2015). Evaluating FQ validity in the CS, Mihura et al. (2013) reported that Conventional (X+%) and Distorted (X-%) Form variables were significantly related to external criteria such as DSM diagnoses or observer ratings (respectively, $r = .48$, $p < .001$, and $r = .49$, $p < .001$) and that X-% appropriately differentiated patients with psychosis from other patients with distorted perceptions (e.g., borderline and schizotypal PD). As for the R-PAS FQ scores, Su and colleagues (2015) reported on the incremental validity of the R-PAS FQ-% and variables to which the FQ codes are crucial subcomponents (i.e., TP-Comp and EII-3) over the CS counterpart (i.e., X-%, PTI, and EII-2) suggesting that improvements in the R-PAS FQ tables have enhanced the interpretive validity of the FQ codings.

Despite the good to excellent support for FQo and FQ-, different studies have shown lower inter-rater reliabilities for FQu codes compared to the other codes. Considering CS variables, Acklin et al. (2000) reported moderate reliabilities at response level for nonpatient ($\kappa = .521$) and clinical ($\kappa = .585$) protocols respectively, whereas kappas for FQo and FQ- were higher than .70 for both nonclinical and clinical protocols. Moreover, at a protocol level of analysis, intraclass correlation coefficients (ICC) for Xu% were poor (ICC = .156) for nonpatient protocols and fair (ICC = .483) for clinical protocols. Meyer et al. (2002) also reported lower, although excellent, reliability values for FQxu (ICC = .93) compared to FQxo (ICC = .98) and FQx- (ICC = .96). Considering R-PAS variables at protocol level, Viglione

et al. (2012) found that R-PAS FQu% showed good reliability (ICC = .64) but lower than reliabilities of FQo% (ICC = .84) and FQ-% (ICC = .81). Recently, Pignolo et al. (2017) reported an excellent reliability for the FQo% (ICC = .82), and fair reliability values for both the FQu% (ICC = .59) and FQ-% (ICC = .53). As for response-level reliabilities, Kivisalu, Lewey, Shaffer, and Canfield (2016, 2017) reported the same pattern, with a lower value for FQu ($\kappa$ = .59) than reliabilities of FQo ($\kappa$ = .77) and FQ- ($\kappa$ = .62). Consistently, Lewey et al. (2019), examining response-level, inter-rater reliability between coders who had only R-PAS training and coders who had both CS and R-PAS training, found that the poorest interrater reliability coefficients were for FQu (R-PAS group: AC = .63, $\kappa$ = .53; CS & R-PAS group: AC = .72, $\kappa$ = .62). Thus, it seems that higher inter-rater reliabilities has been reached for FQo codes, followed by FQ- codes, and that raters had more difficulties to code FQu objects reliably.

From the teaching experience and from previous studies, one of the difficulties with which students struggle the most is coding FQ when objects are not listed in the FQ Tables (Viglione et al., 2017). To reduce examiner's errors, the R-PAS Manual (Meyer et al., 2011) provides a step-by-step method to code FQ. The *Preliminary Step* involves reviewing the response location in the FQ Tables to match the response object in its entirety. If the object is not found, examiners should extrapolate the object's FQ going through the following steps. First, examiners should search objects with *Like Shapes* in the same area (*Step 1*) or the same object in *Like Areas* (*Step 2*), and then examiners should look at subcomponents of objects (*Step 3*). At *Step 4* examiners should *Review the Accumulated Information to Make an FQ Judgement*. Although the R-PAS Manual strongly suggests giving more weight to evidence from the aforementioned steps, examiner's judgements may be made carelessly or with errors.

Although many studies have investigated the validity and inter-rater reliability of FQ codings from both response- and protocol-level perspectives, a major question remains unsolved: how reliably can individuals make FQ judgments in the absence of the FQ tables? The answer to this question has implications for individual examiner's ratings of FQ with individual records. As such, the aim of the present study was to shed light on the ability of Rorschach examiners to code FQ in the absence of the FQ tables and to evaluate the extent to which they agree with each other in evaluating the FQ and FA of response objects, in the absence of the FQ tables. The results of this study may help understanding how examiners would code the FQ when objects are not listed in the FQ Tables.

## Method

### Raters

Because our aim was to evaluate the extent to which examiners could code FQ correctly in the absence of the FQ tables, 21 graduate students in the authors research labs (i.e., research collaborators) from the U.S. and Italy served as raters. All raters were trained in R-PAS coding and had completed at least one semester of Rorschach instruction. Thus, they were well-acquainted with FQ determination and because they had not been exposed to CS coding, were not affected by previous scoring systems or systematic errors in the coding of FQ. All ratings were completed in English. This manuscript should be considered to be an inter-rater reliability lab exercise among researchers in training. Because objects rated by the raters were listed in the FQ Tables and were not taken from responses given by human participants, there are no ethical aspects to disclose.

### The survey

The survey was developed to investigate how raters rated both the FA and FQ of objects without using the FQ tables. As for the FA ratings, we replicated the procedure used by R-PAS authors in developing the R-PAS FQ tables. Raters examined the fit and provided

FA ratings for a list of objects and gave an evaluation based on a five-point Likert scale, that is: 1 = No. I can't see it at all. Clearly, it's a distortion; 2 = Not really. I don't really see that. Overall, it does not match the blot area; 3 = A little. If I work at it, I can sort of see that; 4 = Yes. I can see that. It matches the blot pretty well; 5 = Definitely. I think it looks exactly or almost exactly like that. To identify thresholds to divide FA values in categories that reflected the traditional FQ categories (i.e., -, u, and o), we applied the same cut scores as reported in the R-PAS manual, so that objects with a mean rating of 2.4 or less were evaluated as FA-; objects with a mean rating between 2.5 and 3.4 were evaluated as FAu; objects with a mean rating of 3.5 or more were evaluated as FAo. Moreover, among the three categories of FQ (i.e., FQ-, FQu, and FQo), some objects seem to be more easily classified into each FQ categories than others, so that it is possible to distinguish prototype objects from objects that are considered on the threshold between two categories (Meyer et al., 2011). The division between prototype and threshold objects were made by referring to FA values. FA values lower than 1.75 indicate FQ- prototypes, FA values between 2.85 and 3.05 indicate FQu prototypes, and FA values higher than 4.15 indicate FQo prototypes. As for the thresholds, FA values between 1.90 and 2.20 indicate threshold objects between FQ- and FQu, whereas FA values between 2.55 and 2.75 indicate threshold objects between FQu and FQo.

For the Rorschach, one administers five black and grey cards, two black, grey, and red cards, and three multi-colored cards. Consistent with this relative frequency, we selected two black and grey cards (I and VI), one black and red card (III), and one multi-colored card for our survey. Within each card, we selected commonly used, individual locations because they provide enough FQ table entries to populate the prototype and threshold FA values noted in the paragraph above. Indeed, variability in the number of FQ entries across cards for prototypes and thresholds is due to fluctuations in frequencies across locations in the FQ Tables. Accordingly, uncommon details (Dd) were not used. Since the frequency per record

of whole (W, mean = 9.6) and common details (D, mean = 10.7) are about equal we included two W and two D locations for each card.

Thus, we selected four locations from four different cards: W Location for Cards I and VIII, D1 for Card VI, and D2 for card III. In selecting the entries from the FQ tables, we divided them in prototypes and thresholds according to the R-PAS Manual (Meyer et al., 2011). Prototypes had FA < 1.75 for FQ- (e.g., Card I, W, Bear, FA = 1.55), FA between 2.85 and 3.05 for FQu (e.g., Card III, D2, Hook, FA = 2.95), and FA > 4.15 for FQo (e.g., Card I, W, Insect or Bug (Winged), FA = 4.17). We established two different thresholds between FQ- and FQu: the first threshold for FQ- had FA values between 1.90 and 2.20 (e.g., Card VIII, W, Jacket, FA = 1.98), whereas the second threshold for FQu had FA values between 2.55 and 2.75 (e.g., Card VI, D1, Urn, FA = 2.61). Then, we randomly selected 86 entries that fell within the ranges indicated from the R-PAS Manual (Table 1): 23 response objects for both Card I (W) and Card VIII (W), and 20 objects for both Card III (D2) and Card VI (D1).

[Enter Table 1 about here]

Raters were asked to look at the relevant Rorschach card and response location and rate the fit of each object according to the 5-point FA scale used by the authors of R-PAS in developing the FQ Tables, knowing that, generally, FA values of 1 and 2 represent FQ- codes, an FA value of 3 corresponds to FQu codes, whereas FA values of 4 and 5 are considered FQo. They were also asked to decide on whether they would code FQo, FQu, or FQ-, knowing that 10% of the objects should be coded FQo, about 45% FQu, and about 45% FQ-. The raters could not use the FQ tables, so they rated each entry relying only on their ability to see the objects.

**Data analysis**

In the first step, we considered the FQ classifications made by the raters without looking at the FQ tables in the manual. Because we selected entries from the FQ tables, we were able to determine the degree of convergence between raters' classifications and the R-PAS FQ Tables. In other words, we examined how individual examiners would code a specific entry when left to rely only on their ability to see the objects. Thus, to evaluate whether the raters coded each entry listed in the survey correctly, we computed correct classification and Cohen's kappa values comparing the codes of the raters with those reported in the R-PAS FQ Tables. For Cohen's kappa values, we considered the following cut-offs: kappas between .20 and .40 fair, kappas between .41 and .60 moderate, kappas between .61 and .80 good, and kappas above .80 very good (Altman, 1991; Landis & Koch, 1977).

Second, given that in the development of the FQ Tables FA ratings were used to evaluate the degree to which each object fits with the contour of the inkblot, we examined average FA ratings produced by the raters. We were particularly interested in evaluating whether raters would be able to agree with each other on the degree of fit to the inkblot of the selected entries. To do that, we computed two-way random Intraclass Correlations (ICC) between average FA values by the raters and those used by R-PAS authors in developing the R-PAS FQ Tables. For ICC values, we considered the following cut-offs: ICCs < .40 poor reliability, ICCs between .40 and .59 fair reliability, ICCs between .60 and .74 good reliability, and ICCs of .75 or above excellent reliability (Cicchetti, 1994; Shrout & Fliess, 1979).

## Results

Correct classifications consisted of the percentage of correct FQ classifications of all the 86 entries by the 21 raters (Table 2). The overall hit rate was 58.5% and the percentage of correct classification was low for FQu objects (46.2%), higher for FQ- entries (68.1%), and

the highest for FQo entries (74.3%). Correct classifications of each Card closely reflect the overall correct classification (Table 2). With regard to each Card, Card VI obtained the lower overall hit rate (51.2%), whereas the highest value was obtained for Card III (61.6%). As for the highest FQ classification by Card, 71.3% of FQ- entries from Card I and Card VII were recognized by the raters, 50.6% of FQu entries were correctly classified in Card VIII, and 85.4% of FQo entries were correctly classified in Card III. However, less than 50% of FQu entries from Card I, III, and VI were correctly classified by the raters. In general, Cohen's kappa was fair ($\kappa = .338$), ranging from .200 for Card VI to .392 for Card III.

As for Prototype and Threshold entries, Table 2 shows that hit rates of Prototype entries were generally higher than of Threshold entries. The overall correct classification for Prototype entries was 62.9%, with 75.8% for FQ- entries, 48.7% for FQu entries, and 74.3% for FQo entries, whereas the correct classification for Threshold objects was 53.2%, with 61.8% of FQ- entries and 43.2% of FQu entries being correctly classified. Interestingly, considering Prototypes, 23 FQ- entries were classified as FQo. The most misclassified FQ- entry was "Skeleton" in Card VIII (W Location), followed by "Bug" in Card VI (D1 Location), which were coded FQo by four and three raters respectively. On the other hand, three FQo Prototype entries were classified as FQ- by the raters. The entry that was mostly misclassified was "Flower" in Card VIII (W Location), which was coded FQ- by seven raters. Cohen's kappa for Prototypes was moderate ($\kappa = .439$), ranging from .277 for Card VI to .488 for Card I, whereas Cohen's kappa for Thresholds was poor ($\kappa = .156$), ranging from .032 for Card I to .233 for Card III.

[Enter Table 2 about here]

To analyze the fit (i.e., Form Accuracy) of the entries, we asked the raters to rate each object on the 5-point scale, where 1 indicated a poor fit and 5 indicated an optimum fit. According to the R-PAS Manual, if one were to rely only on FA/Fit, objects with a FA rating of 2.4 or less would be classified as FQ-, objects with a FA of 3.5 or above would be classified as FQo, and objects with FA between 2.4 and 3.5 would be classified as FQu. As would be expected, mean FA ratings ($M$ = 1.99, $SD$ = .95) related to FQ- entries were lower than the suggested threshold of 2.4 and the mean value of FQo entries was higher than 3.5 ($M$ = 4.17, $SD$ = 1.02), whereas the mean FQu rating ($M$ = 2.96, $SD$ = 1.00) was in the intervening range (Table 3). Considering Prototypes and Thresholds, FQo Prototypes should have a mean FA above 4.15, FQu Prototypes a mean FA between 2.85 and 3.05, whereas FQ- Prototype should have a mean FA lower than 1.75. As shown in Table 3, FQ- and FQu Prototypes had a mean FA higher than the cut-off, with mean FA values of 1.81 ($SD$ = .94) and of 3.18 ($SD$ = .92), respectively. This pattern is consistent for Cards III and VI, whereas for Card I and VIII the mean FA ratings of FQ- Prototypes were lower than 1.75. On the other hand, FQo Prototypes showed mean FA ratings higher than 4.15, with the exception of Card VIII ($M$ = 3.83, $SD$ = 1.27). As for Thresholds, (Table 3) FA mean ratings were between the suggested range for both FQ- and FQu Thresholds. However, FA mean ratings for FQu Thresholds were lower than 2.55 for Card I ($M$ = 2.35, $SD$ = 1.00) and higher than 2.75 for Card VI ($M$ = 3.06, $SD$ = 1.10).

[Enter Table 3 about here]

To compare the FA ratings by the raters with those used to develop the R-PAS FQ Tables, we computed ICCs. Considering all the entries, the ICC value was .850, indicating an excellent reliability (Cicchetti, 1994; Shrout & Fliess, 1979). However, when looking at the

different FQ codes, ICC coefficients were .403 for FQ- entries, .377 for FQu entries, and, surprisingly, .146 for FQo entries. The unexpected results for FQo entries, lead us to an in-depth analysis of the FA mean values for FQo entries. We found that one entry (i.e., "Flower (Can include leaf)" in Card VIII, Location W) had a FA mean value ($M$ = 2.86, $SD$ = 1.01) lower than 3.5, the cut-off used for FQo categories. Thus, excluding this entry from the analysis, the ICC value for FQo objects became .593. Thus, the results may suggest that, on aggregate ratings, raters were capable of recognizing the fit of the objects to the contour of the inkblot.

## Discussion and Conclusion

The present study evaluated the extent to which Rorschach examiners agree with each other in evaluating the FQ and FA of response objects, in the absence of the FQ tables. The aim was to understand how examiners would code the FQ when objects are not listed in the FQ Tables, and, thus, to investigate examiner's judgements. We asked 21 raters to rate FA and FQ for 86 objects from Card I, III, VI, and VIII. Considering FQ codes, the overall hit rate was 58.5% and the percentages of correct classification were 68.1% for FQ- objects, 46.2% for FQu objects, and 74.3% for FQo objects. The results indicate that examiner judgements are not reliable, and coders should not rely on their opinion in coding FQ but should use all the evidence gathered from the steps listed in the R-PAS Manual in coding FQ. On the other hand, considering FA values, the ICC value was .850, indicating an excellent reliability. Thus, examiner judgements for FQ are inaccurate, but they seem more accurate when they have to establish the degree to which an entry fit the contour of the inkblot. In other words, ICC values indicate that the raters evaluated FA of each entry consistently with the raters who evaluated FA for the R-PAS FQ Tables.

From the results of the present study, two main implications are worth noting. First, these findings may shed light on the lower inter-rater reliability values related to FQu

compared to FQ- and FQo. One may speculate that when examiners are forced to make individual judgements in coding FQ because the object is not listed in the FQ tables (Step 4 of the instruction given by the R-PAS Manual), they would produce inconsistent codings. In this direction, future studies should evaluate potential differences in the inter-rater reliability values for the FQ codes between FQ classifications based on the Manual (Step 1 to 3) and FQ classifications based on individual judgements (Step 4). Second, in terms of training, particular attention should be paid to the steps described in the manual on how to code FQ when the object is not listed in the Manual. New learners who found the coding of FQ particularly difficult (Viglione et al., 2017) may find some comfort in knowing all the strategies they should adopt to deal with this challenge.

Although this study is the first to analyze the impact of examiners' judgements on the coding of FQ, some limitations are worth noting. First, we administered the survey to a small sample of graduate student collaborators. Expert researchers and clinicians may thus yield higher levels of reliability with the FQ tables, considering the experience they may have accumulated in coding objects not listed in the FQ tables. However, given that most of the studies evaluating the inter-rater reliability of Rorschach scores are based on the codings made by graduate students or young researchers and clinicians, we believe that our findings reflect the real context in which these studies were conducted. Second, we selected only single objects in W and D location, and we did not consider multiple objects or Dd locations. Given that our aim was to evaluate the extent to which raters would be able to code FQ variables correctly without using the FQ tables, we decided to maintain stable the level of difficulty of the coding. Indeed, coding one object in one location is easier that coding multiple objects in multiple or uncommon location.

## References

Acklin, M. W., McDowell, C. J., Verschell, M- S., & Chan, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment, 74*(1), 15-47. http://dx.doi.org/10.1207/S15327752JPA740103

Altman, D. G. (1991). *Practical statistics for medical research.* London: Chapman and Hall.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. http://dx.doi.org/10.1037/1040-3590.6.4.284

Exner, J. E., Jr. (1974). *The Rorschach: A Comprehensive System: Vol. 2. Interpretation*. New York, NY: Wiley.

Exner, J. E., Jr. (2003). *The Rorschach: A Comprehensive System: Vol. 1. Basic foundations and principles of interpretation* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Kivisalu, T. M., Lewey, J. H., Shaffer, T. W., & Canfield, M. L. (2016). An Investigation of Interrater Reliability for the Rorschach Performance Assessment System (R–PAS) in a Nonpatient U.S. Sample. *Journal of Personality Assessment, 98*(4), 382-390. http://dx.doi.org/10.1080/00223891.2015.1118380

Kivisalu, T. M., Lewey, J. H., Shaffer, T. W., & Canfield, M. L. (2017). Correction to: An Investigation of Interrater Reliability for the Rorschach Performance Assessment System (R–PAS) in a Nonpatient U.S. Sample. *Journal of Personality Assessment, 99*(5), 558-560. http://dx.doi.org/10.1080/00223891.2017.1325244

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-74. http://dx.doi.org/10.2307/2529310

Lewey, J. H., Kivisalu, T. M., & Giromini, L. (2019). Coding With R-PAS: Does Prior Training with the Exner Comprehensive System Impact Interrater Reliability

Compared to Those Examiners with Only R-PAS-Based Training? *Journal of Personality Assessment, 101*(4), 393-401. http://dx.doi.org/10.1080/00223891.2018.1476361

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Fowler, J. C., Piers, C. C., & Resnick, J. (2002). An examination of interrater reliability for scoring the Rorschach Comprehensive System in eight data sets. *Journal of Personality Assessment, 78*(2), 219-274. http://dx.doi.org/10.1207/S15327752JPA7802_03

Meyer, G. J., Viglione, D. J., Mihura, J. L., Erard, R. E., & Erdberg, P. (2011). *Rorschach performance assessment system: Administration, coding, interpretation, and technical manual*. Toledo, OH: Rorschach Performance Assessment System.

Mihura, J. L., Meyer, G. J., Dumitrascu, N., & Bombel, G. (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the Comprehensive System. *Psychological Bulletin, 139*(3), 548-605. http://dx.doi.org/10.1037/a0029406

Pignolo, C., Giromini, L., Ando', A., Ghirardello, D., Di Girolamo, M., Ales, F., & Zennaro, A. (2017). An Interrater Reliability Study of Rorschach Performance Assessment System (R–PAS) Raw and Complexity-Adjusted Scores. *Journal of Personality Assessment, 99*(6), 619–625. http://dx.doi.org/10.1080/00223891.2017.1296844

Rorschach, H. (1921). *Psychodiagnostik Talfen*. Bern: Huber.

Shrout, P. E., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Su, W. S., Viglione, D.J., Green, E. E., Tam, W.C., Su, J.A., & Chang, Y. T. (2015) Cultural and linguistic adaptability of the Rorschach Performance Assessment System as a measure of psychotic characteristics and severity of mental disturbance in Taiwan. *Psychological Assessment, 27*(4), 1273-85. http://dx.doi.org/10.1037/pas0000144

Viglione, D. J., Meyer, G. J., Resende, A. C., & Pignolo, C. (2017). A Survey of Challenges Experienced by New Learners Coding the Rorschach. *Journal of Personality Assessment, 99*(3), 315–323. http://dx.doi.org/10.1080/00223891.2016.1233559

Viglione, D.J, Blume-Marcovici, A. C., Miller, H., L., Giromini, L., & Meyer, G. (2012). An Inter-Rater Reliability Study for the Rorschach Performance Assessment System. *Journal of Personality Assessment, 94*(6), 607–612. http://dx.doi.org/10.1080/00223891.2012.684118

**Table 1**

*FQ Entries listed in the survey*

| | Card | | | | Total |
|---|---|---|---|---|---|
| | I | III | VI | VIII | |
| FQo | 4 | 2 | 0 | 2 | 8 |
| Prototypes | 4 | 2 | 0 | 2 | 8 |
| FQu | 10 | 9 | 10 | 11 | 40 |
| Prototypes | 6 | 5 | 5 | 6 | 22 |
| Thresholds | 4 | 4 | 5 | 5 | 18 |
| FQ- | 9 | 9 | 10 | 10 | 38 |
| Prototypes | 3 | 5 | 5 | 4 | 17 |
| Thresholds | 6 | 4 | 5 | 6 | 21 |
| Total | 23 | 20 | 20 | 23 | 86 |
| Prototypes | 13 | 11 | 10 | 12 | 47 |
| Thresholds | 10 | 8 | 10 | 11 | 39 |

**Table 2**

*FQ correct classifications and Cohen's kappa between raters and R-PAS Manual*

| | | Overall | | Card I | | Card III | | Card VI | | Card VIII | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | # of ratings | CC% | κ | CC% | κ | CC% | κ | CC% | κ | CC% | κ |
| FQ- | 798 | 68.1 | | 71.3 | | 69.8 | | 60.4 | | 71.3 | |
| FQu | 840 | 46.2 | .338 | 43.8 | .375 | 48.1 | .392 | 42.0 | .200 | 50.6 | .366 |
| FQo | 168 | 74.3 | | 72.6 | | 85.4 | | - | | 66.7 | |
| Total | 1806 | 58.5 | | 59.6 | | 61.6 | | 51.2 | | 61.0 | |
| Prototype | | | | | | | | | | | |
| FQ- | 357 | 75.8 | | 85.7 | | 73.3 | | 69.5 | | 79.5 | |
| FQu | 462 | 48.7 | .439 | 51.6 | .488 | 45.2 | .454 | 43.3 | .277 | 53.2 | .454 |
| FQo | 168 | 74.3 | | 72.6 | | 85.4 | | - | | 66.7 | |
| Total | 987 | 62.9 | | 66.0 | | 63.6 | | 56.5 | | 64.2 | |
| Threshold | | | | | | | | | | | |
| FQ- | 441 | 61.8 | | 64.0 | | 65.5 | | 51.0 | | 65.9 | |
| FQu | 378 | 43.2 | .156 | 32.1 | .032 | 51.8 | .233 | 40.8 | .122 | 47.6 | .217 |
| Total | 819 | 53.2 | | 51.2 | | 58.7 | | 45.9 | | 57.6 | |

*Note.* CC% = Correctly Classified %: refers to the % of ratings that identified the correct FQ level.

**Table 3**

*Descriptive statistics of FA ratings and ICCs (R = 1806)*

| | # of objects | FQ- | | FQu | | FQo | | *ICC* |
|---|---|---|---|---|---|---|---|---|
| | | *M* | *DS* | *M* | *DS* | *M* | *DS* | |
| Overall | 86 | 1.99 | 0.95 | 2.96 | 1.00 | 4.17 | 1.02 | .850 |
| Prototype | 47 | 1.81 | 0.94 | 3.18 | 0.92 | 4.17 | 1.02 | .884 |
| Theshold | 39 | 2.14 | 0.92 | 2.70 | 1.04 | - | - | .538 |
| Card I | 23 | 1.94 | 0.91 | 2.86 | 1.04 | 4.21 | 0.97 | .939 |
| Prototype | 13 | 1.54 | 0.86 | 3.21 | 0.91 | 4.21 | 0.97 | .943 |
| Theshold | 10 | 2.14 | 0.88 | 2.35 | 1.00 | - | - | .492 |
| Card III | 19 | 1.97 | 0.92 | 3.04 | 1.01 | 4.41 | 0.71 | .892 |
| Prototype | 11 | 1.87 | 0.96 | 3.37 | 0.94 | 4.41 | 0.71 | .906 |
| Theshold | 8 | 2.10 | 0.84 | 2.63 | 0.96 | - | - | .690 |
| Card VI | 20 | 2.17 | 1.01 | 3.10 | 1.04 | - | - | .673 |
| Prototype | 10 | 1.98 | 0.96 | 3.13 | 0.99 | - | - | .795 |
| Theshold | 10 | 2.37 | 1.02 | 3.06 | 1.10 | - | - | .415 |
| Card VIII | 23 | 1.88 | 0.92 | 2.87 | 0.92 | 3.83 | 1.27 | .828 |
| Prototype | 12 | 1.73 | 0.93 | 3.02 | 0.85 | 3.83 | 1.27 | .850 |
| Theshold | 11 | 1.98 | 0.91 | 2.70 | 0.96 | - | - | .656 |

**Summary**

Form Quality (FQ) is an essential variable that has been recognized for its importance since the development of the Rorschach Inkblot test. It refers to the "goodness of fit" of visualized objects to the corresponding area of the blot used by the examinee. Moreover, FQ scores are a well-validated measure of perception accuracy, reality testing, and severity of psychological disturbance. Research studies reveal that inter-rater reliability of FQ scoring is good when visualized objects are available in FQ tables. However, many visualized objects are not found in the FQ tables so that scoring must rely on one's individual judgment. No research has directly asked the question of how reliably and accurately can individuals make these FQ judgments in the absence of the FQ tables. If the answer were to be "not very good" then such difficulty would limit the validity of FQ scoring and a remedy might be in order. To address this question about examiner judgment of fit in terms of FQ scoring accuracy and inter-rater reliability, we used the Rorschach Performance Assessment System (R-PAS) method. We asked 21 graduate students (i.e., research collaborators) from our research labs to rate Form Accuracy (FA) and FQ for 86 objects from a subset of four Rorschach card (I, III, VI, and VIII). The results clearly reveal that individual examiner making FA judgements without using the FQ tables are not reliable. These findings shed light on the lower inter-rater reliability values related to FQu compared to FQ- and FQo. When scoring FQ, one should carefully scrutinize the empirically support of the FQ tables and base the FQ score on these rather than personal judgement. For R-PAS there are procedures to follow in the manual and online in an effort to maximize accuracy and reliability. In terms of training, new learners who found the coding of FQ particularly difficult may find some comfort in knowing all the strategies they should adopt to deal with this challenge.

**Riassunto**

La Qualità Formale (Form Quality; FQ) è una variabile essenziale che è stata riconosciuta per la sua importanza sin dallo sviluppo del test di Rorschach. Si riferisce alla "bontà dell'adattamento" degli oggetti visualizzati all'area della macchia utilizzata dall'esaminato. Inoltre, i punteggi FQ sono una misura validata di accuratezza della percezione, dell'esame di realtà, e della gravità del disturbo psicologico. Diversi studi hanno rivelato che l'affidabilità tra giudici delle codifiche FQ sia buona quando gli oggetti visualizzati sono elencati nella tabella FQ. Tuttavia, molti oggetti visualizzati non sono presenti nelle tabelle FQ cosicché lo scoring deve fare affidamento sul giudizio individuale del clinico. Nessuna ricerca ha indagato direttamente quanto affidabili e accurati siano i giudizi individuali sulle codifiche FQ in assenza delle tabelle FQ. Se la risposta dovesse essere "non molto" allora questa difficoltà limiterebbe la validità dello scoring di FQ. Per affrontare questo problema sul grado di accuratezza e affidabilità tra giudici dei giudizi degli esaminatori nel siglare FQ abbiamo utilizzato il metodo Rorschach Performance Assessment System (R-PAS). Abbiamo chiesto a 21 dottorandi (collaboratori di ricerca) di valutare l'Accuratezza Formale (Form Accuracy; FA) e di siglare FQ per 86 oggetti delle tavole I, III, VI e VIII. I risultati rivelano chiaramente che i singoli giudizi degli esaminatori nel valutare FA senza l'utilizzo delle tavole FQ non sono affidabili. Questi risultati potrebbero far luce sui valori di affidabilità tra giudici più bassi relativi a FQu rispetto a FQ- e FQo. Quando si sigla FQ, si dovrebbe esaminare attentamente le tavole FQ che derivano da supporto empirico e basare la codifica FQ sulle tavole FQ piuttosto che su giudizi individuali. Nel metodo R-PAS vengono presentate le procedure da seguire sia nel manuale sia online per massimizzare l'accuratezza e l'affidabilità. In termini di training, i nuovi esaminatori che trovano particolarmente difficile

codificare FQ possono trovare conforto nel conoscere tutte le strategie che dovrebbero adottare per affrontare questa sfida.

**Résumé**

La qualité formelle (Form Quality; FQ) est une variable essentielle qui a été reconnue pour son importance depuis le développement du test de Rorschach. Elle fait référence à la "qualité de l'ajustement" des objets visualisés aux contours de la tâche utilisée par le patient. De plus, les scores FQ constituent une mesure bien validée de la précision de la perception, du test de réalité et de la gravité du trouble psychologique. Des études ont révélé que la fiabilité inter-juges des encodages FQ est bonne lorsque les objets affichés sont répertoriés dans le tableau FQ. Cependant, de nombreux objets affichés ne sont pas présents dans les tableaux FQ, de sorte que la notation doit reposer sur le jugement individuel du clinicien. Aucune recherche n'a directement examiné la fiabilité et l'exactitude des jugements individuels sur les codages FQ en l'absence de tableaux FQ. Si la réponse était «pas beaucoup», alors cette difficulté limiterait la validité de la notation FQ. Pour résoudre ce problème du degré d'exactitude et de fiabilité parmi les juges des jugements des examinateurs lors de la signature du FQ, nous avons utilisé la méthode du Rorschach Performance Assessment System (R-PAS). Nous avons demandé à 21 doctorants (collaborateurs de recherche) d'évaluer l'exactitude formelle (Form Accuracy; FA) et FQ pour 86 objets des planches I, III, VI et VIII. Les résultats révèlent clairement que les jugements individuels des examinateurs lors de l'évaluation de la FA sans l'utilisation des tableaux FQ ne sont pas fiables. Ces résultats pourraient expliquer les valeurs de fiabilité des juges les plus faibles concernant FQu comparées à FQ- et FQo. Lors de l'initialisation de FQ, il faut examiner attentivement les tableaux FQ qui découlent d'un soutien empirique et baser le codage FQ sur les tableaux FQ plutôt que sur des jugements individuels. La méthode R-PAS présente les procédures à suivre à la fois dans le manuel et en ligne pour maximiser l'exactitude et la fiabilité. En termes de

formation, les nouveaux examinateurs trouvant qu'il est particulièrement difficile de codifier FQ peuvent être soulagés de connaître toutes les stratégies possibles à adopter pour relever ce défi.

**Resumen**

La calidad de la forma (FQ) es una variable esencial que ha sido reconocida por su importancia desde el desarrollo de la prueba de Rorschach. Se refiere a la "bondad de ajuste" de los objetos expuestos al área de la mancha utilizada por el examinador. Además, los puntajes FQ son una medida validada de la precisión de la percepción, las pruebas de realidad y la gravedad del trastorno psicológico. Varios estudios han revelado que la confiabilidad entre jueces de las codificaciones FQ es buena cuando los objetos mostrados se enumeran en la tabla FQ. Sin embargo, muchos de los objetos mostrados no están presentes en las tablas FQ, por lo que la puntuación debe basarse en el juicio individual del médico. Ninguna investigación ha investigado directamente qué tan confiables y precisos son los juicios individuales sobre la codificación FQ en ausencia de tablas FQ. Si la respuesta fuera "no mucho", esta dificultad limitaría la validez de la puntuación FQ. Para abordar este problema del grado de precisión y confiabilidad entre los jueces de los juicios de los examinadores al firmar FQ, usamos el método Rorschach Performance Assessment System (R-PAS). Solicitamos a 21 estudiantes de doctorado (colaboradores de investigación) que evaluaran la Exactitud de la Forma (Form Accuracy; FA) y firmaran FQ para 86 objetos en las tablas I, III, VI y VIII. Los resultados revelan claramente que los juicios individuales de los examinadores al evaluar AF sin el uso de las tablas FQ no son confiables. Estos resultados podrían arrojar luz sobre valores más bajos de confiabilidad entre jueces para FQu en comparación con FQ- y FQo. Al inicializar FQ, se deben examinar cuidadosamente las tablas de FQ que se derivan del soporte empírico y basar la codificación de FQ en las tablas de FQ en lugar de juicios individuales. El método R-PAS presenta los procedimientos a seguir tanto en el manual como en línea para maximizar la precisión y confiabilidad. En términos de

formación, los nuevos examinadores a los que les resulte particularmente difícil codificar QF pueden encontrar consuelo al conocer todas las estrategias que deben adoptar para afrontar este desafío.